CODE FOR SIMULATING DEGREES OF FREEDOM FOR THE ITEMS IN A PRINCIPAL COMPONENTS ANALYSIS OF VARIANCE

BU-1488 -M **June, 2000**

Walter T. Federer
Dept. of Biometrics
Cornell University

and

Russell D. Wolfinger
SAS Institute, Inc.
R52, SAS Campus Dr.
Cary, NC 27513

Keywords: none provided by authors.

Abstract: Purpose: To provide simulations required to approximate degrees of freedom for such items as principal components, auto-regression, smoothing, kriging, and the like.

Bu 1484
M

Title: Code for simulating degrees of freedom for the items in a principal components analysis of variance.

Authors: Walter T. Federer, 434 Warren Hall, Cornell University, Ithaca, NY 14853; e-mail WTF1@cornell.edu. and Russell D. Wolfinger, SAS Institute, INC. R53, SAS Campus Drive, Cary, NC 27513.

Purpose: To provide simulations required to approximate degrees of freedom for such items as principal components, auto-regressions, smoothing, kriging, and the like.

Data: The simulation is for the data obtained from the plan given below, Table 12.5 of W. G. Cochran and G. M. Cox (1957), Experimental Designs, Wiley, New York. The experiment design is a balanced lattice square with v = 16 insecticide treatments and r = 5 replicates (complete blocks). The measurement y is the average of three counts of plants infected with boll weevil. Grad is the linear polynomial regression coefficients for the column order.
The simulation uses the randomization plan of the experiment for each simulation and unit random normal deviates. The sum of squares for a line in the analysis of variance is an estimate of the degrees of freedom since the expected value of each mean square is one. The randomization plan of the input data is used for each simulation.

```
/* Here the data the data are included but an infile statement may be used to input the plan of the experiment to be
simulated. */
data original;
input y rep row col grad treat;
  label row='incomplete block';
  datalines;
9.0   1 1 1 -3 10
20.3  1 1 2 -1 12
17.7  1 1 3  1  9
26.3  1 1 4  3 11
 4.7  1 2 1 -3  2
 9.0  1 2 2 -1  4
 7.3  1 2 3  1  1
 8.3  1 2 4  3  3
 9.0  1 3 1 -3 14
 6.7  1 3 2 -1 16
11.7  1 3 3  1 13
 4.3  1 3 4  3 15
 4.0  1 4 1 -3  6
 5.0  1 4 2 -1  8
 5.7  1 4 3  1  5
14.3  1 4 4  3  7
19.0  2 1 1 -3  5
 8.7  2 1 2 -1 12
13.0  2 1 3  1 15
15.7  2 1 4  3  2
12.0  2 2 1 -3 10
 6.0  2 2 2 -1  7
15.3  2 2 3  1  4
12.0  2 2 4  3 13
12.7  2 3 1 -3 16
 6.3  2 3 2 -1  1
 1.7  2 3 3  1  6
13.0  2 3 4  3 11
 3.7  2 4 1 -3  3
 3.7  2 4 2 -1 14
 8.0  2 4 3  1  9
13.3  2 4 4  3  8
17.0  3 1 1 -3 10
 7.0  3 1 2 -1 15
```

```
10.3 3 1 3  1  8
 1.3  3 1 4  3  1
11.3 3 2 1 -3  9
12.3 3 2 2 -1 16
 3.0  3 2 3  1  7
 5.3  3 2 4  3  2
12.3 3 3 1 -3 12
 8.7  3 3 2 -1 13
 8.0  3 3 3  1  6
 9.3  3 3 4  3  3
30.3 3 4 1 -3 11
22.3 3 4 2 -1 14
11.0 3 4 3  1  5
12.7 3 4 4  3  4
 5.0  4 1 1 -3 16
10.3 4 1 2 -1 12
 5.7  4 1 3  1  8
12.7 4 1 4  3  4
 2.7  4 2 1 -3 11
 6.7  4 2 2 -1 15
10.3 4 2 3  1  3
 5.7  4 2 4  3  7
 1.0  4 3 1 -3  1
10.3 4 3 2 -1  5
11.3 4 3 3  1  9
11.7 4 3 4  3 13
11.0 4 4 1 -3  6
19.0 4 4 2 -1  2
20.7 4 4 3  1 14
29.7 4 4 4  3 10
 2.0  5 1 1 -3  3
 5.0  5 1 2 -1 16
 4.0  5 1 3  1  5
13.7 5 1 4  3 10
 9.3  5 2 1 -3  6
 1.7  5 2 2 -1  9
 6.3  5 2 3  1  4
12.3 5 2 4  3 15
16.7 5 3 1 -3 12
 4.3  5 3 2 -1  7
18.7 5 3 3  1 14
 8.7  5 3 4  3  1
16.7 5 4 1 -3 13
30.0 5 4 2 -1  2
25.7 5 4 3  1 11
14.0 5 4 4  3  8
run;

/* data sets for the pc analysis */
proc sort data=original;
  by rep row col;
run;

%let nsim=2;  /* nsim=2 is for 2 simulations.  Usually nsim will be large. */
%let seed=2834701;  /* Any random seed may be specified. */
```

```
data sim;
  set original;
  do k=1 to &nsim;
    y = rannor(&seed); /* This statement says that unit normal random deviates are to be used in the simulation. */
    output;
  end;
run;

/* principal component analysis, by k, i. e. for each simulated analysis, and rep */
proc sort data=sim;
  by k rep col row;
proc transpose data=sim prefix=row out=simr(drop=_name_);
  by k rep col;
  var y;
proc princomp data=simr prefix=rpc n=2 out=rowvar noprint;
  by k rep;
  var row1-row4; /* Four rows in the design. */
proc sort data=sim;
  by k rep row col;
proc transpose data=sim prefix=col out=simc(drop=_name_);
  by k rep row;
  var y;
proc princomp data=simc prefix=cpc n=2 out=colvar noprint;
  by k rep;
  var col1-col4; /* Four columns in the design. */

/* expand data sets and merge */
data cc;
  set colvar;
  array colv{4} col1-col4;
  do col = 1 to 4;
    y = colv{col};
    output;
  end;
  drop col1-col4;
data rr;
  set rowvar;
  array rowv{4} row1-row4;
  do row = 1 to 4;
    y = rowv{row};
    output;
  end;
  drop row1-row4;
proc sort data=rr;
  by k rep row col;
data ana;
  merge sim cc rr;
  by k rep row col;

/* analysis of variance using the principal components, non-nested */
proc glm data=ana outstat=o1 noprint;
  by k;
  class rep treat;
  model y=rep treat cpc1 cpc2 rpc1 rpc2 cpc1*rpc1 cpc1*rpc2
    cpc2*rpc1 cpc2*rpc2 ;
proc print data=o1;
```

```
run;

/* using the principal components, nested */
proc glm data=ana outstat=o2 noprint;
  by k;
  class rep treat;
  model y=rep treat cpc1(rep) cpc2(rep) rpc1(rep) rpc2(rep)
    cpc1*rpc1(rep) cpc1*rpc2(rep) cpc2*rpc1(rep) cpc2*rpc2(rep) ;
proc print data=o2;
run;

/* using the textbook analysis of the design as in Cochran and Cox (1957), page 493, and as given above. This
provides a check on the simulations as the sums of squares are the degrees of freedom. */
proc glm data=ana outstat=o3 noprint;
  by k;
  class rep row col treat;
  model y=rep treat row(rep) col(rep);
  lsmean treat;
proc print data=o3;
run;
```

The output of the above program is given below. SS1 is type I sum of squares, SS3 is type III sum of squares, and the sum of squares is the degrees of freedom as the expected value of each mean square in the table is one.

Unnested PCTA ANOVA - run 1

OBS	K	_NAME_	_SOURCE_	_TYPE_	DF	SS	F	PROB
1	1	Y	ERROR	ERROR	52	41.0582	.	.
2	1	Y	REP	SS1	4	10.6315	3.3662	0.01594
3	1	Y	TREAT	SS1	15	16.8648	1.4239	0.17144
4	1	Y	CPC1	SS1	1	9.5108	12.0454	0.00105
5	1	Y	CPC2	SS1	1	0.0890	0.1127	0.73848
6	1	Y	RPC1	SS1	1	5.9200	7.4976	0.00844
7	1	Y	RPC2	SS1	1	1.1131	1.4097	0.24049
8	1	Y	CPC1*RPC1	SS1	1	0.4601	0.5828	0.44869
9	1	Y	CPC1*RPC2	SS1	1	2.9147	3.6915	0.06018
10	1	Y	CPC2*RPC1	SS1	1	0.0440	0.0558	0.81427
11	1	Y	CPC2*RPC2	SS1	1	0.0642	0.0813	0.77664
12	1	Y	REP	SS3	4	10.6315	3.3662	0.01594
13	1	Y	TREAT	SS3	15	5.6731	0.4790	0.94084
14	1	Y	CPC1	SS3	1	9.0891	11.5113	0.00133
15	1	Y	CPC2	SS3	1	0.3867	0.4898	0.48715
16	1	Y	RPC1	SS3	1	6.1529	7.79260	0.00732
17	1	Y	RPC2	SS3	1	0.9345	1.18358	0.28165
18	1	Y	CPC1*RPC1	SS3	1	0.6489	0.82188	0.36881
19	1	Y	CPC1*RPC2	SS3	1	2.9887	3.78517	0.05712
20	1	Y	CPC2*RPC1	SS3	1	0.0470	0.05948	0.80827
21	1	Y	CPC2*RPC2	SS3	1	0.0642	0.08133	0.77664

Unnested PCTA ANOVA - run 2

22	2	Y	ERROR	ERROR	52	38.1947	.	.
23	2	Y	REP	SS1	4	1.9787	0.67346	0.61338
24	2	Y	TREAT	SS1	15	19.1134	1.73479	0.07252
25	2	Y	CPC1	SS1	1	1.7968	2.44618	0.12388
26	2	Y	CPC2	SS1	1	1.8594	2.53141	0.11766
27	2	Y	RPC1	SS1	1	6.0489	8.23524	0.00593
28	2	Y	RPC2	SS1	1	1.1192	1.52375	0.22260

29	2	Y	CPC1*RPC1	SS1	1	5.9688	8.12616	0.00624
30	2	Y	CPC1*RPC2	SS1	1	0.0026	0.00354	0.95277
31	2	Y	CPC2*RPC1	SS1	1	0.09327	0.12699	0.72302
32	2	Y	CPC2*RPC2	SS1	1	0.74168	1.00975	0.31962
33	2	Y	REP	SS3	4	1.97867	0.67346	0.61338
34	2	Y	TREAT	SS3	15	4.41711	0.40091	0.97267
35	2	Y	CPC1	SS3	1	3.54125	4.82123	0.03260
36	2	Y	CPC2	SS3	1	1.62114	2.20710	0.14341
37	2	Y	RPC1	SS3	1	6.56799	8.94197	0.00425
38	2	Y	RPC2	SS3	1	1.24960	1.70127	0.19787
39	2	Y	CPC1*RPC1	SS3	1	5.87599	7.99984	0.00663
40	2	Y	CPC1*RPC2	SS3	1	0.01225	0.01668	0.89773
41	2	Y	CPC2*RPC1	SS3	1	0.06838	0.09310	0.76149
42	2	Y	CPC2*RPC2	SS3	1	0.74168	1.00975	0.31962

Nested PCTA ANOVA - run 1

OBS	K	_NAME_	_SOURCE_	_TYPE_	DF	SS	F	PROB
1	1	Y	ERROR	ERROR	20	4.7282	.	.
2	1	Y	REP	SS1	4	10.6315	11.2427	0.00006
3	1	Y	TREAT	SS1	15	16.8648	4.7558	0.00076
4	1	Y	CPC1(REP)	SS1	5	15.5141	13.1248	0.00001
5	1	Y	CPC2(REP)	SS1	5	2.7152	2.2970	0.08381
6	1	Y	RPC1(REP)	SS1	5	10.0044	8.4637	0.00020
7	1	Y	RPC2(REP)	SS1	5	4.7658	4.0318	0.01080
8	1	Y	CPC1*RPC1(REP)	SS1	5	7.5124	6.3554	0.00110
9	1	Y	CPC1*RPC2(REP)	SS1	5	4.6495	3.9335	0.01203
10	1	Y	CPC2*RPC1(REP)	SS1	5	8.2553	6.9839	0.00064
11	1	Y	CPC2*RPC2(REP)	SS1	5	3.0294	2.5628	0.06004
12	1	Y	REP	SS3	4	10.6315	11.2427	0.00006
13	1	Y	TREAT	SS3	15	3.1950	0.9010	0.57497
14	1	Y	CPC1(REP)	SS3	5	13.1371	11.1139	0.00003
15	1	Y	CPC2(REP)	SS3	5	2.6241	2.2200	0.09243
16	1	Y	RPC1(REP)	SS3	5	8.5463	7.2301	0.00052
17	1	Y	RPC2(REP)	SS3	5	4.3279	3.6614	0.01630
18	1	Y	CPC1*RPC1(REP)	SS3	5	3.5137	2.9726	0.03638
19	1	Y	CPC1*RPC2(REP)	SS3	5	3.0198	2.5547	0.06065
20	1	Y	CPC2*RPC1(REP)	SS3	5	5.8240	4.9271	0.00423
21	1	Y	CPC2*RPC2(REP)	SS3	5	3.0294	2.5628	0.06004

Nested PCTA ANOVA - run 2

22	2	Y	ERROR	ERROR	20	4.3221	.	.
23	2	Y	REP	SS1	4	1.9787	2.2890	0.09551
24	2	Y	TREAT	SS1	15	19.1134	5.8963	0.00018
25	2	Y	CPC1(REP)	SS1	5	9.1962	8.5108	0.00019
26	2	Y	CPC2(REP)	SS1	5	3.9972	3.6993	0.01562
27	2	Y	RPC1(REP)	SS1	5	8.7444	8.0927	0.00026
28	2	Y	RPC2(REP)	SS1	5	0.8850	0.8190	0.55043
29	2	Y	CPC1*RPC1(REP)	SS1	5	15.2570	14.1200	0.00001
30	2	Y	CPC1*RPC2(REP)	SS1	5	6.2267	5.7626	0.00188
31	2	Y	CPC2*RPC1(REP)	SS1	5	3.81879	3.53420	0.01883
32	2	Y	CPC2*RPC2(REP)	SS1	5	3.37785	3.12611	0.03029
33	2	Y	REP	SS3	4	1.97867	2.28901	0.09551
34	2	Y	TREAT	SS3	15	1.79833	0.55477	0.87591
35	2	Y	CPC1(REP)	SS3	5	6.68211	6.18412	0.00128
36	2	Y	CPC2(REP)	SS3	5	3.52047	3.25810	0.02592
37	2	Y	RPC1(REP)	SS3	5	8.15456	7.54683	0.00040
38	2	Y	RPC2(REP)	SS3	5	1.92512	1.78165	0.16248

39	2	Y	CPC1*RPC1(REP)	SS3	5	8.04745	7.44771	0.00043
40	2	Y	CPC1*RPC2(REP)	SS3	5	7.22510	6.68665	0.00082
41	2	Y	CPC2*RPC1(REP)	SS3	5	4.65856	4.31137	0.00799
42	2	Y	CPC2*RPC2(REP)	SS3	5	3.37785	3.12611	0.03029

Textbook ANOVA - run 1

OBS	K	_NAME_	_SOURCE_	_TYPE_	DF	SS	F	PROB
1	1	Y	ERROR	ERROR	30	25.6177	.	.
2	1	Y	REP	SS1	4	10.6315	3.11254	0.02958
3	1	Y	TREAT	SS1	15	16.8648	1.31665	0.25255
4	1	Y	ROW(REP)	SS1	15	19.6950	1.53760	0.15374
5	1	Y	COL(REP)	SS1	15	15.8615	1.23832	0.29886
6	1	Y	REP	SS3	4	10.6315	3.11254	0.02958
7	1	Y	TREAT	SS3	15	9.8877	0.77194	0.69613
8	1	Y	ROW(REP)	SS3	15	15.7929	1.23297	0.30227
9	1	Y	COL(REP)	SS3	15	15.8615	1.23832	0.29886

Textbook ANOVA - run 2

10	2	Y	ERROR	ERROR	30	29.5471	.	.
11	2	Y	REP	SS1	4	1.9787	0.50225	0.73428
12	2	Y	TREAT	SS1	15	19.1134	1.29376	0.26542
13	2	Y	ROW(REP)	SS1	15	16.4413	1.11289	0.38676
14	2	Y	COL(REP)	SS1	15	9.8368	0.66584	0.79576
15	2	Y	REP	SS3	4	1.9787	0.50225	0.73428
16	2	Y	TREAT	SS3	15	9.3513	0.63298	0.82440
17	2	Y	ROW(REP)	SS3	15	13.8467	0.93726	0.53691
18	2	Y	COL(REP)	SS3	15	9.8368	0.66584	0.79576